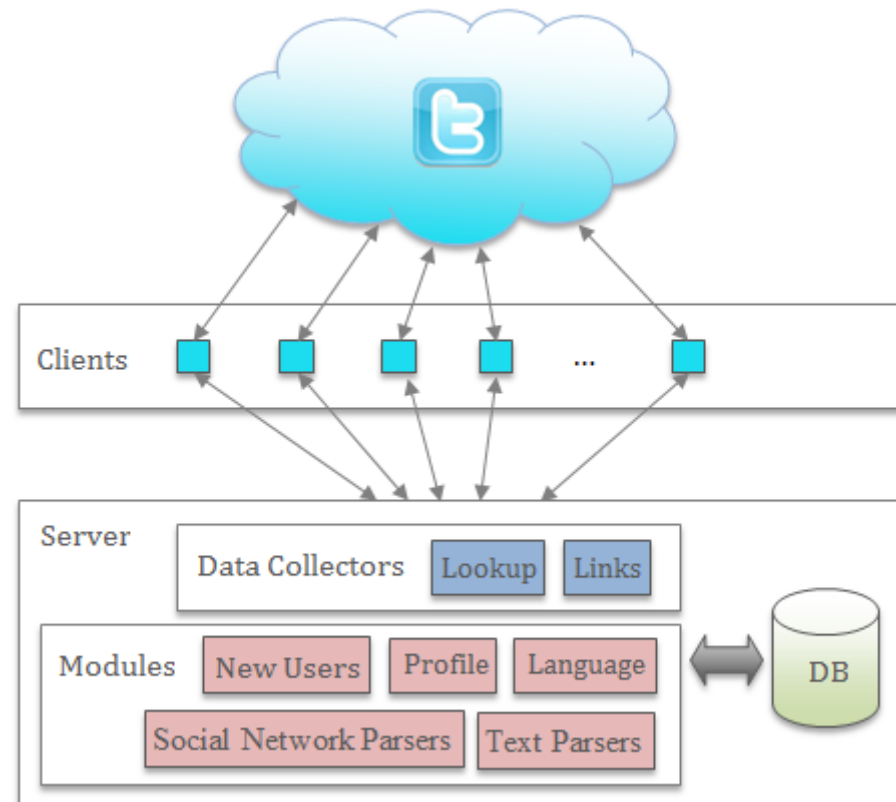


TwitterEcho crawler

Eduardo Oliveira, José Martins, Eduarda
Mendes Rodrigues, Luís Sarmento,
Jorge Teixeira, Matko Bošnjak

The architecture

- Distributed system
- Server
 - Coordinating download process
- Client
 - Executing the download through API



Core functionality

- Downloading
 - Tweets
 - User info
 - Statistics
 - Friends & followers
- Filtering
 - Language (portuguese)
 - Verifying users
 - Tracking active and inactive users

How does it work?

- Start with 2k of users
- Download their tweets and monitor their activity
- Download their friends and followers
- For each new friend, follower or mentioned user, assess their “Portugueseness”
- Add Portuguese users to the seed
- Repeat

Modules

- Language identification
 - Distinguishes Portuguese from Brazilian Portuguese
- User filtering
 - Filtering only Portuguese users
- Social network entity extraction
 - Explicit graphs
 - Friends & followers
 - Implicit graphs
 - Replies, mentions, retweets, hashtags, URLs

Data statistics

- 10.07.: > 5GB of data
 - 75k users
 - 4M tweets
 - 250k of followers lists
 - 150k of friends lists
 - 6M of statistics entry

